Rebecca Hernandez-Gerber, Emily Nabasny, Pamela Vizner
Digital Preservation
October 29, 2013

Archive-it: The Strong and Play Partners

Introduction

For our Archive-It collection we chose to work with staff from The Strong to archive their website, and subsequent related sites. The Strong - located in Rochester, New York - is an umbrella organization dedicated to the study of play in all its forms. This larger institution is comprised of several smaller groups, referred to as Play Partners, including the National Museum of Play, the International Center for the History of Electronic Games, the National Toy Hall of Fame, the Brian Sutton-Smith Library & Archives of Play, and the American Journal of Play.

Our group wanted to archive the website of The Strong because of our previous work with them during the Pressing Restart: Community Discussions on Video Game Preservation conference. Originally, we reached out to JP Dyson, Director of the International Center for the History of Electronic Games, as a potential curator for this project. He passed us onto Allison McGrath, who is The Strong Director of Online and Graphic Media Services, for further details, and it was Allison with whom we developed a plan for this project. After reaching out to JP and Allison about the project we were informed that the institution will be creating a new web architecture at the start of the new year, which made this the perfect opportunity to archive the old sites and thus they happily accepted to work with us on this project.

Project Planning

Perhaps the most difficult part of this project was in deciding exactly which of the related websites to crawl. Due to the independent nature of all Play Partners under the larger institution of The Strong, there are actually six independent websites that operate together. Archiving the websites of the umbrella organization would necessitate archiving six large seed websites, a task that is difficult to consider even under the best of circumstances and funding. As such, the project became an exercise in how to structure a crawl in such a way that limiting the scope does not limit essential parts of the website's materials. It also was an important lesson in how communication with a client, and a presentation of options, is necessary when creating a crawl.

Originally, Allison asked that we start by archiving only the main website for The Strong, and the site for The Museum of Play. She did not request that we archive any of the other Play Partner organizations, even though they technically fall under The Strong heading. We were concerned that our client may have assumed that all Play Partners would automatically be archived if The Strong website linked to them, without knowing that crawling scopes are determined by the initial seeds used. To ensure that we were all looking at this crawl from the same perspective, we rephrased the question and asked our client if she specifically wanted to

crawl the other Play Partner websites and listed each one. Asking the question in a different way resulted in a different answer, with our client now requesting that we archive those websites, as well. It is possible that she simply changed her mind but it is also possible that listing technical details to a client without archival training results in a better understanding of the purpose of a crawl. In the future, this would be a lesson to keep in mind when working with non-archival clients, especially when describing complex archival terms via email to people whose job keeps them busy enough to ignore long and complicated messages.

We also realized that our client did not request any social media to be archived, perhaps ignoring this was possible. We proposed the inclusion of The Strong's Facebook page in the crawls in order to archive how this institution interacts with visitors and Allison agreed.

Project Proposal

The web archiving proposal we developed for our client included sections on project background, guidelines for the curator, project scope, intellectual property and privacy concerns, metadata requirements, and a description of the final product we would produce. Our curator & client, Allison McGrath, agreed to be responsible for directing the development and execution of this plan by determining the scope of what was to be archived as well as deciding on the frequency of the crawls. Additionally, she agreed to notify us of any intellectual property and privacy concerns. In response to this last point, our client stated that all images and recordings on the website were owned by The Strong under a Creative Commons License, which was added to the proposal's section on intellectual property and privacy concerns.

All other portions of the proposal were left to the group to decide upon, due to the lack of direct statements and communications from our client. We agreed to begin with several test crawls, before moving on to one active crawl. These crawls would have a frequency of once a month and a length of seventy-two hours (this last one set by default on Archive-it for monthly crawls). The scope was determined to be the seven seed URLs that constitute all Play Partner websites listed above as well as the Facebook page managed by The Strong and National Museum of Play.

We also consulted with our client regarding metadata requirements for the collection. They stated that they did not have any specific requirements, therefore we suggested to use some basic fields in order to make the collection searchable on the Archive-it website. Of the fourteen fields Archive-it offers as initial metadata we used title, creator, subject, description and publisher. However, we agreed to add more fields if they became necessary.

Our final product consisted of a number of steps. First, we agreed to discuss any final adjustments with our client curator. Once the crawl was accessible, we also asked that the client curator access it and suggest further modifications before summarizing and evaluating the project.

Crawling

Test Crawl Nr. 1:

Our first test crawl began with six seed URLs. One for The Strong main site, and one for each of the sites for The Play Partners. As a result of this being our first test, we selected a time frame of one day in order to test the number of documents and the size of the collection. The crawl stopped before completing due to the short time limit, and we realized that one day was not enough to archive even the basic features of each website. In this crawl 35,369 documents were captured.

Most of the documents that were considered out of scope were .css and js files as well as youtube videos, which we initially did not intend on including in the scope. We also found that the test crawl only saved one video, from The Museum of Play site. At this stage we did not use the Scope-it crawl explorer because we already wanted to expand the crawl.

Test Crawl Nr. 2/Other Test Crawls:

For our second test we changed the time limit of our six seed URLs. This time we had the crawl run for three days in order to expand the scope of the previous crawl, which had stopped because of time limit. The original seed websites were archived correctly, including 86,607 documents and fifteen videos. The same .css and .js files were again out of scope, which we found were very important later on.

We included the Facebook page after initiating the second test, due to delays in the response from our curator, Allison. Thus, in order to evaluate the crawl of this site we ran a separate test only for this Facebook page. Unfortunately, we realized only after the crawl that the Facebook page was blocked from being archived by a robots.txt. It took us time to realize that Facebook required the extra step of modifying the crawl scope to ignore robots.txt.

Official Crawl

For our official crawl we chose to run a "one time crawl" for one day to have a quick evidence of what the final crawl would look like; before performing a "monthly crawl" as determined initially in our agreement with the institution.

We included the six websites used as seed URLs in the previous test plus the Facebook page. This time, we checked the box "ignore robots.txt" to allow the crawl to archive Facebook. However, these was not done until after the crawl started, so Facebook page was not archived in this step. We finally realized that any changes made after the crawl started will not affect current crawls. In order to archive the Facebook site, we ran a 1-hour crawl only with this seed, which was successful.

The final monthly crawl ran for 3 days after which we finally had access to the archived content. Browsing the archived website we noticed that the websites were archived properly, except for the database collections available online. This is when we realized that to be able to see this content correctly we would need the .css and .js files. Browsing the help documentation we found that the crawler normally has issues with these files, but we had to contact the Archive-

it team to look for a solution. According to them, this is what they call a " crawler trap", which creates some kind of loop of links, where the crawler just leaves this files as "out of scope". In order to solve this, we had to add a host constraint for our six websites, which will block any URL from being archived if this trap occurs. At the time of writing, we were running a patch crawl to incorporate this changes.

Following this we created QA reports for the crawls and realized that Facebook, as well as some of the other seeds, reported QA issues, URLs blocked by robots.txt, and capture issues. Subsequently we ran patch crawls, but the issues remained. We believe that those issues have something to do with the .css or .js files, which changes will hopefully be incorporated after the last patch crawl we scheduled.

Final Results with Video Analysis

All websites were successfully archived, except for the databases of the collections online, which were considered out of scope because of the "crawler traps". At the time of writing, we do not know if the patch crawls with the recommended changes will allow the databases to be archived.

Facebook presented a special problem in our project, mostly because we added it late and because it was not very clear how to deactivate the robots blocking the site. However, once we figured out how to handle it, we had no further issues.

We ultimately archived 33 videos, in formats .mp4 and .flv. We did not have any problems watching this videos using the WayBack Machine, although browsing the .mp4 files using the timeline or jumping to a specific spot was not possible. All videos from YouTube were bypassed in the initial crawls, all of them were .flv. We could finally archive these videos during one of the final crawls, although their content was clearly out of scope.

This issue could present a special problem for institutions that have YouTube embedded content in their websites. The immediate solution for this case, if the institution wants to archive them, would be to archive those separately, using "crawl one page only" feature or as a separate seed on the collection, considering also all the difficulties involved in archiving YouTube pages. Archive-it also offers the option to archive only video files, ignoring the typical layout of the site and any other special feature. This could be a solution for institutions especially interested in the content leaving the looks of YouTube behind.

We finally decided not to use Scope-it since the number of documents archived and the size of the collection were appropriate. The major problems we had were related to social media sites, which we chose to fix that manually.

Something worth mentioning is that Archive-it does not provide clear information about the videos that were not archived or out of scope, unless you start digging in the long list of "out of scope" URLs for each host, which limited our evaluation regarding this matter.

Evaluation & Feedback

Many of our issues when dealing with this project have stemmed from a lack of direction and communication with our client. While we do not expect that the client will have a comprehensive understand of the technical details of crawling or to direct us in those areas, it is necessary that they evaluate the data we send them and suggest further modifications to our work. It is difficult to proceed without this information, and it has been sadly lacking. Significant delays in response over time contribute to this effect, especially considering the limited amount of time given to complete this project, leaving us with no other choice but to proceed under the assumption that a lack of communication indicates assent to the work done so far. Were this a professional client, we could not be held responsible for satisfying their demands, as we would often be unsure as to what those demands are in the first place. In that sense we believe that communicating this type of information over email clearly worsened this delay.

Moving forward with this project, we will adjust the future crawls to reflect any modifications that JP or Allison request, as well as any changes that we find necessary for a more complete and accurate archive of the websites. We hope that many of the problems we have now will be solved after our last patch crawl.